Available on CMS information server

CMS DP -2023/010

The Compact Muon Solenoid Experiment

CMS Performance Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

22 March 2023

An AutoEncoder-based Anomaly Detection tool with a per-LS granularity

CMS Collaboration

Abstract

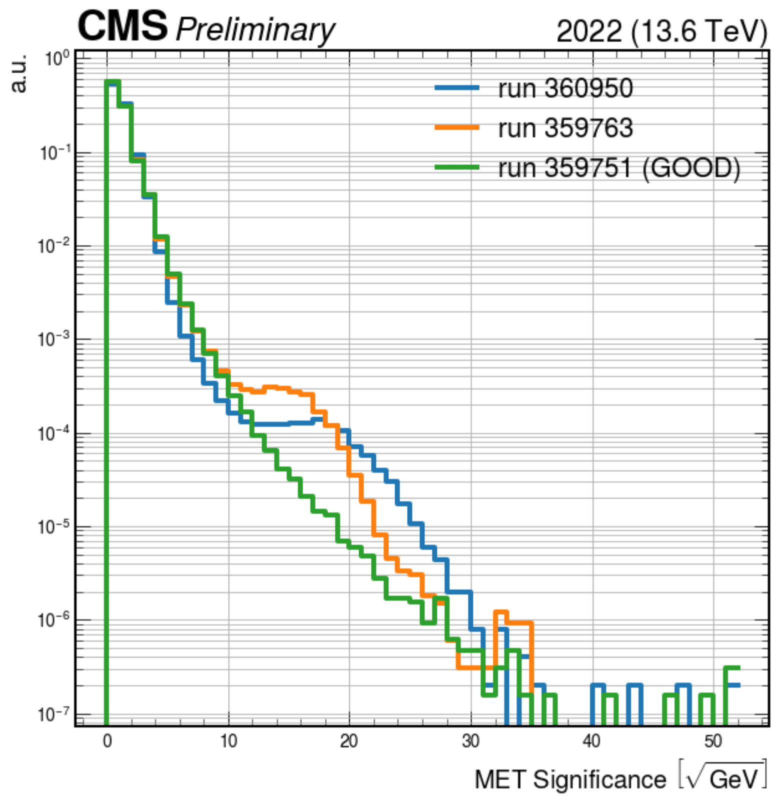
An AutoEncoder-based Anomaly Detection Tool capable of detecting anomalies in DQM Monitor Elements with a per-Lumisection granularity is presented.

An AutoEncoder-based Anomaly Detection tool with a per-LS granularity

CMS Collaboration

cms-phys-conveners-JME@cern.ch

Introduction



Histograms of a Monitor Element (MET Significance) for three different runs chosen as example for this note, one flagged GOOD and two presenting an anomaly, therefore flagged BAD.

- In CMS, Data Certification (DC) is the final step of quality checks performed by *Data Quality monitoring* (DQM) on recorded collision events.
- Data is gathered in luminosity sections, lumisections in short (LSs), corresponding to ~23 seconds of data taking.
- LSs are grouped in runs. For each run, experts monitor a number of reconstructed distributions called Monitor Elements (MEs) to spot issues in the data.
- For the specific case of quantities pertaining to hadronic jets and missing transverse momentum (MET), an issue in a few LSs would cause the entire run to be flagged as problematic (*BAD*), and thus removed from the pool of "good-for-analysis" data (*GOOD*).

Per-LS data

- In CMS, the possibility of accumulating quantities monitored for data quality purposes **per-LS** has been recently extended to *Jet and Missing Energy* (JME) MEs.
- This possibility allows for a higher granularity detection of anomalies, potentially enabling the saving of higher amounts of data from runs presenting only a limited set of anomalous LSs. Given the high number, $O(1000)$, of LSs to be analyzed for each run, an **automated approach** (rather than a manual one) for DC is required.
- Machine Learning (ML), particularly Neural Networks (NN), can be implemented to this end.
- An unsupervised ML model based on a specific NN architecture called AutoEncoder (AE) is employed.

AutoEncoder-based Anomaly Detection Tool

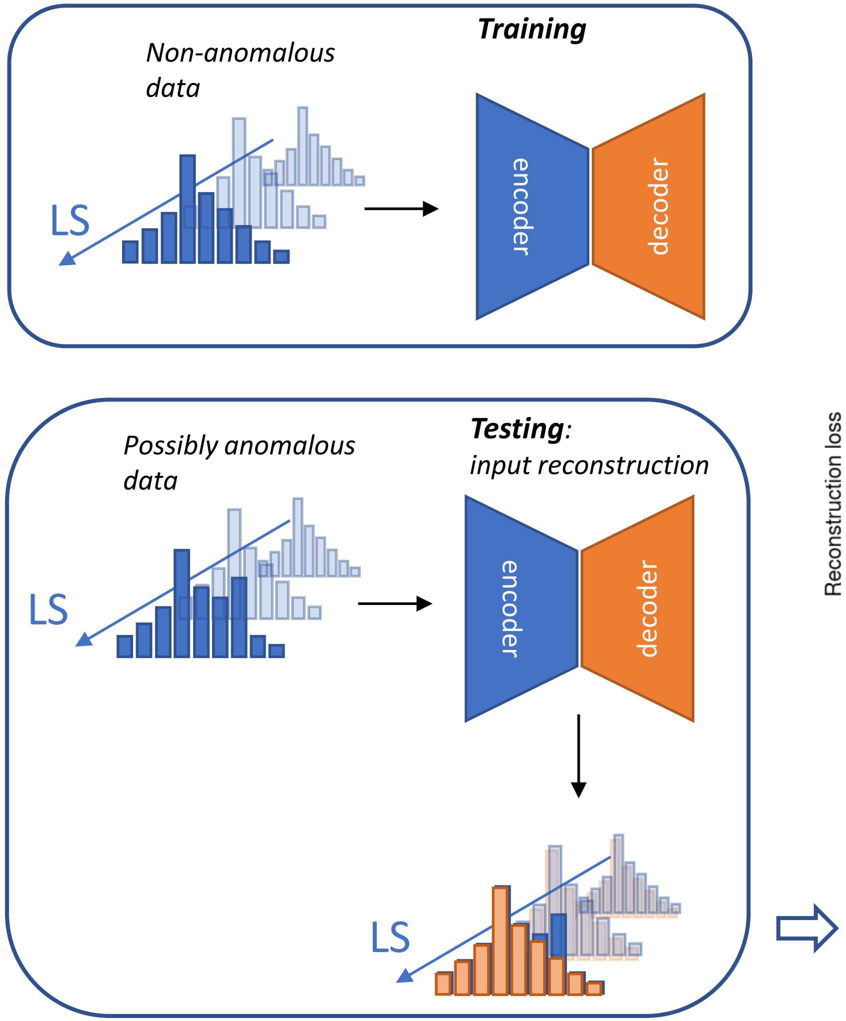
- The model is trained on non-anomalous data from *GOOD* runs: histograms of specific MEs are fed to the model with an LS granularity to allow the AE to learn a «normal» non-anomalous behavior of that specific ME. The training is performed via the minimization of the reconstruction loss, a measure of the distance between the input and output of the AE. In this case the reconstruction loss is the mean squared error:

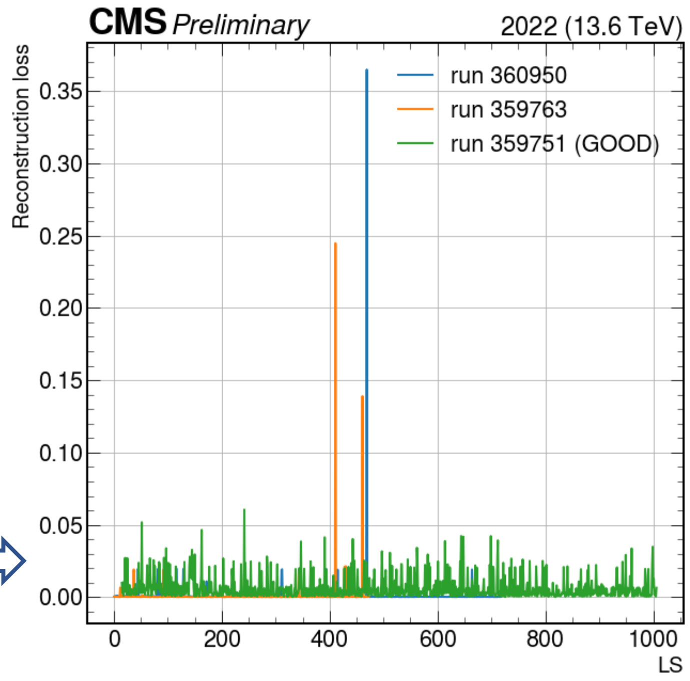
$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \hat{y}_i)^2$$

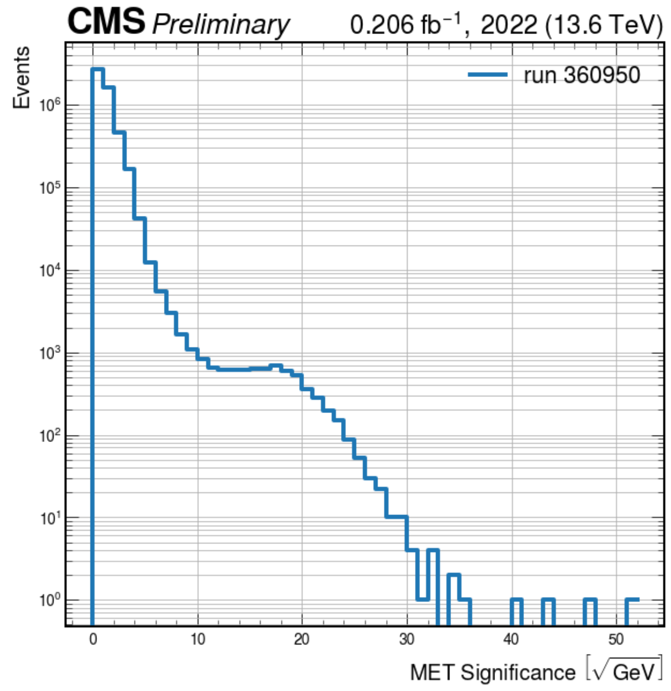
where y and $\hat{y}$ are respectively the input and the output of the AE and n is the bin number.

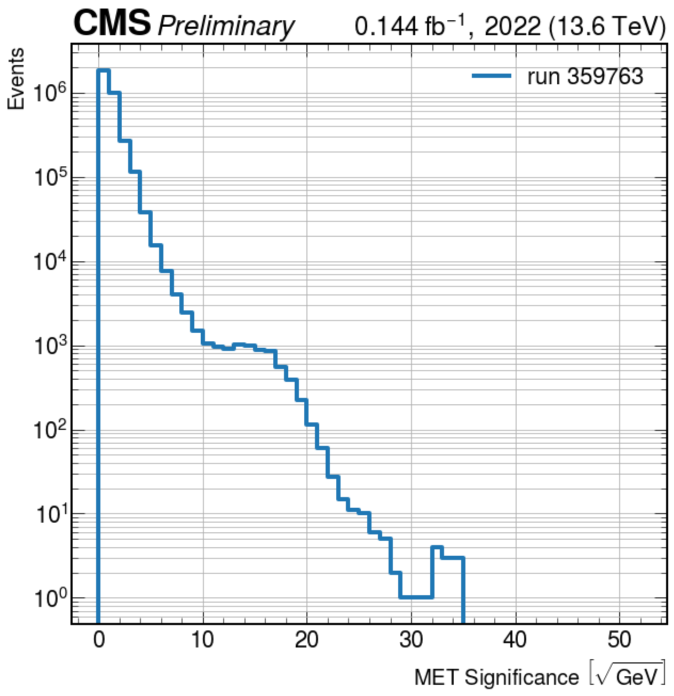
- Possibly anomalous runs under investigation are tested by looking again at the reconstruction loss: peaks in this function indicate LSs containing histograms that deviate from the learned behavior.

- The comparison between the reconstruction losses of the three runs under study is on the right.







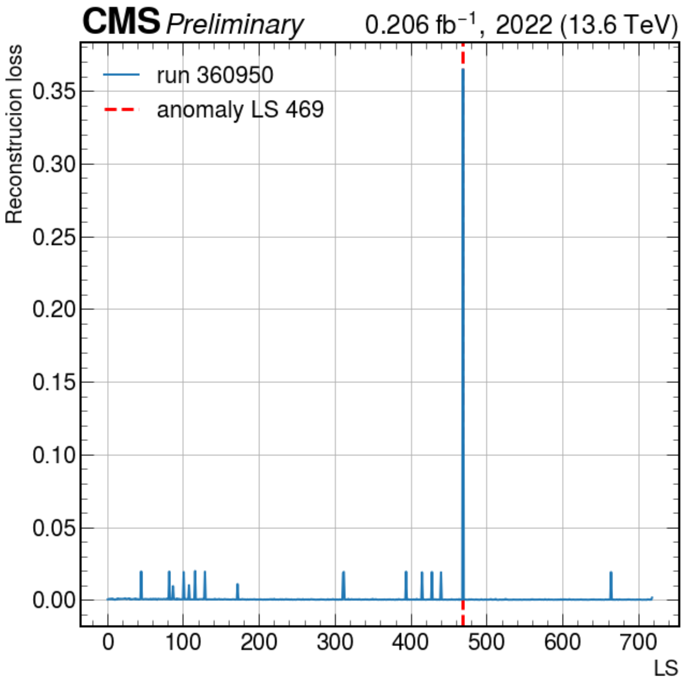


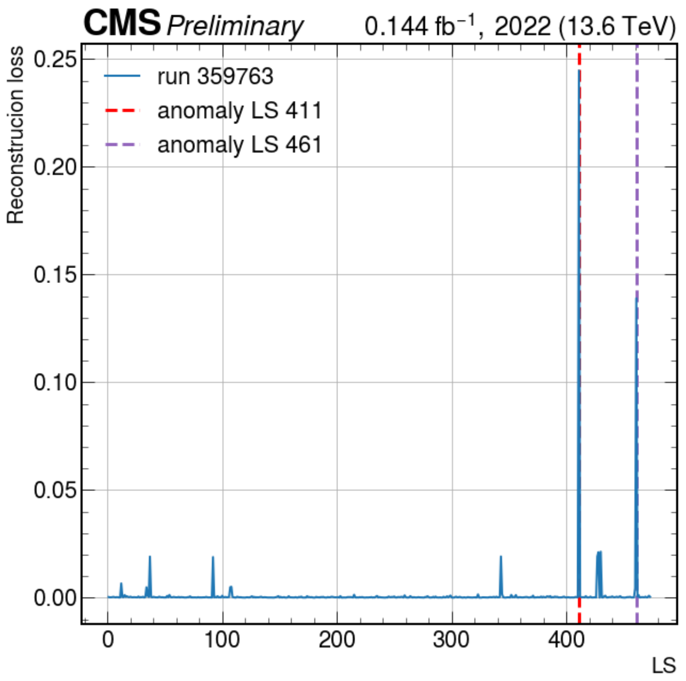
Runs 360950 & 359763

- Both runs show a visible bump on many MEs, one of them is *MET Significance*.
- *MET Significance:*

$$METSig \equiv \frac{MET}{\sqrt{SumET}} = \frac{MET}{\sqrt{\sum|\vec{p}_T|}}$$

- Both runs were flagged *BAD* by JME DQM.

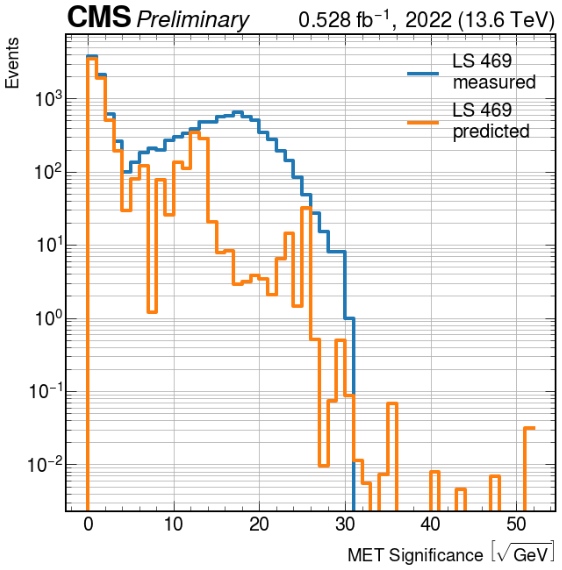




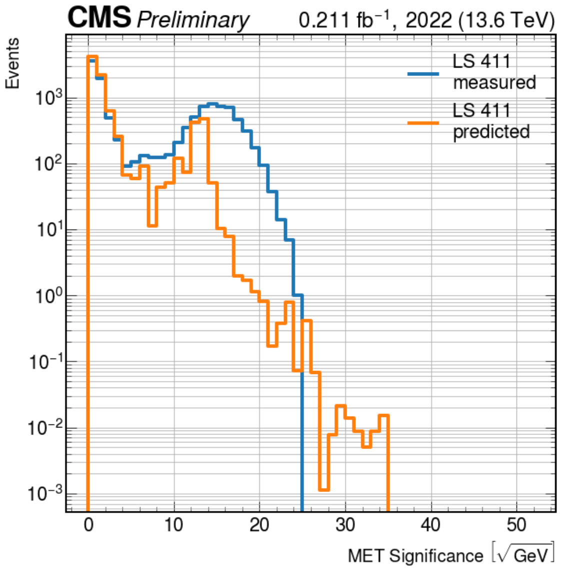
Identifying the anomalies

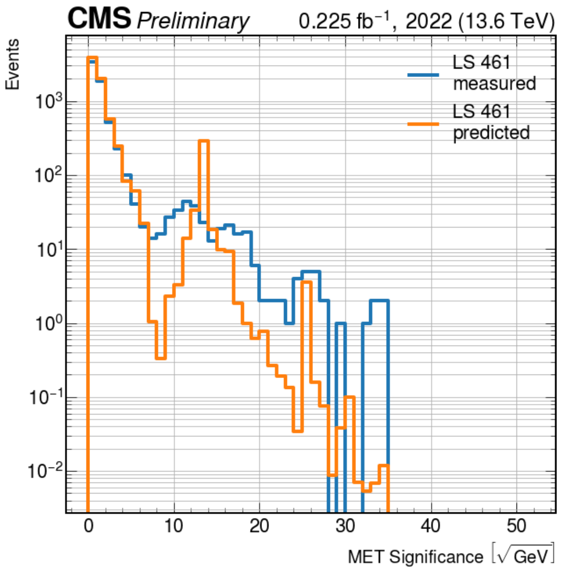
- By analyzing the per-LS MET Significance for both runs via the AE-based anomaly detection tool, we found peaks in the reconstruction loss limited to a small number of LSs.
- Run 360950 presents a peak corresponding to LS 469.
- Run 359763 presents two peaks, the biggest one corresponding to LS 411, the smaller one corresponding to LS 461.

Input vs prediction



Run 360950 input and prediction for LS 469.

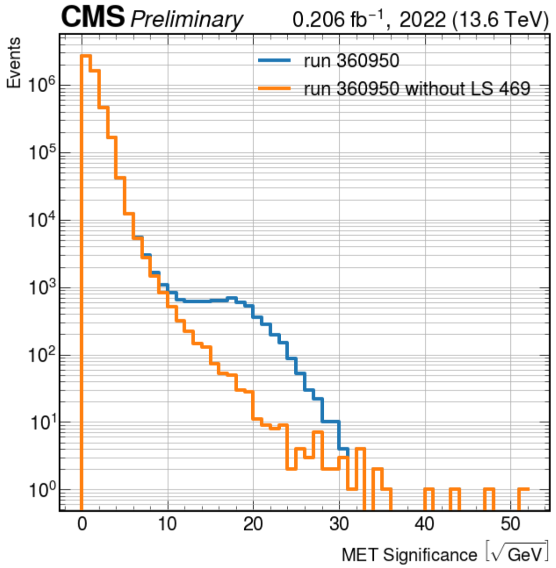




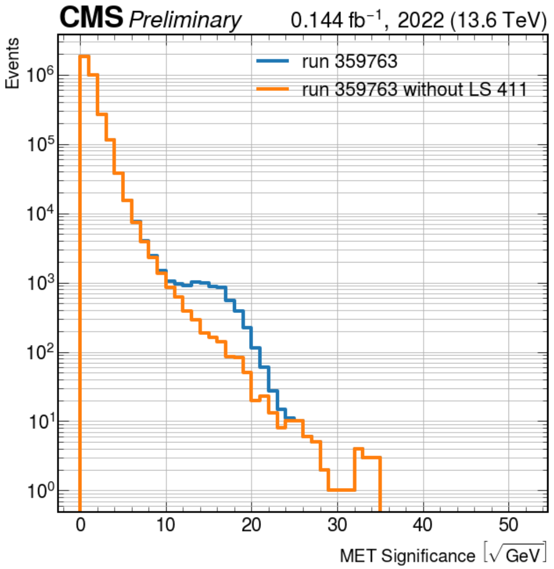
Run 359763 input and prediction for LS 411 (left) and for LS 461 (right).

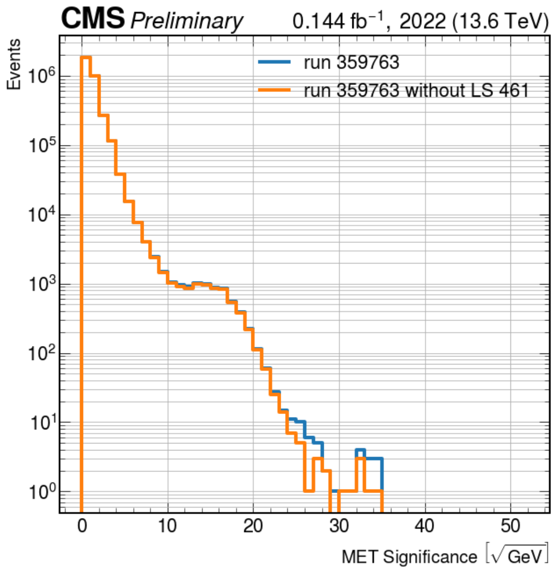
- What's causing the peaks in the reconstruction loss is the difference between the input of the AE and the output (prediction).
- On the top the histogram of LS 469 of run 360950.
- On the bottom left the histogram of LS 411 of run 359763.
- On the bottom right the histogram of LS 461 of run 359763.

Both runs: removing the anomalies



Run 360950 with and without LS 469.





Run 359763 with and without LS 411 (left) and with and without LS 461 (right).

- Once anomalous LSs are identified they are removed from the run.
- The resulting histograms for both *BAD* runs show how the cause of the MET Significance bump was LS 469 for run 360950 and LS 411 for run 359763.
- The removal of LS 461 smooths out the tail of the histogram.

Conclusions

- We developed an AutoEncoder-based Anomaly Detection Tool capable of detecting anomalies in DQM MEs with a per-LS granularity.
- We tested the tool on several runs flagged *BAD* by JME DQM and identified the source of the anomalous behavior in a limited set of LSs.
- In particular, we removed one LS from each run presented in this note and verified that the remainder was no more anomalous.
- The equivalent luminosity recovered from the two runs is $\sim 350\ pb^{-1}$.
- Exploiting the per-LS granularity in DQM and systematically employing the tool we presented will enable an increase in efficiency of the DC procedure, ultimately resulting in a larger dataset available to physics analyses.